Example-based Machine Translation

Cristina Vertan
University of Hamburg • Informatics Department
Natural Language Systems Group

WWW: http://nats-www.informatik.uni-hamburg.de/~cri/
E-Mail: vertan@informatik.uni-hamburg.de

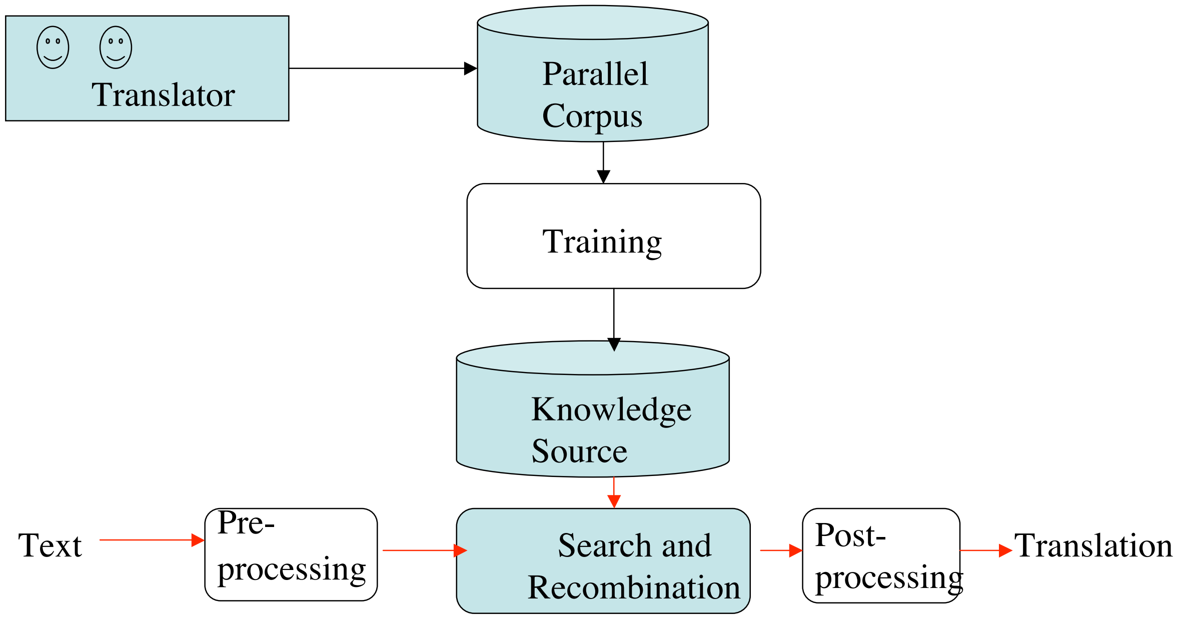
Contents

- General Principles of Corpus-based Machine Translation ⇦
- Principles and functionality of EBMT Systems
- Adding morphological and syntactic knowledge to EBMT systems
- Adding semantic knowledge to EBMT systems

General Principles -Corpus based MT

- The linguistic phenomena in both languages as well as the transfer rules are no longer linguistically described but derived automatically from a parallel corpus.
- First an aligned corpus is built
- Next step is a training phase, in which are calculated the connections between elements in the source language as well as in the target language (sometimes the results are called „knowledge sources“)).
- The translation is the result of 2 processes:
  - A search process (of elements in the source language)
  - A best-evaluated relation with a target expression
- There are 2 types of corpus-based MT systems
  - Example based MT - The translation of a source text is based of translation examples in the database
  - Statistical MT - the alignment information from the corpus is used for the training of a statistical translation model

Generic Architecture of a corpus-based MT-system



Aligned Corpus

- A parallel Corpus:
  - Is a collection of texts in at least 2 languages. It is extremely important that the content is the same for both texts.
  - Examples: Official Documents from EU , Newspapers in Countries with more than 1 official language
  - Contains markers (tags) for content-identical elements (Sentences, Paragraphs) in Texts
- The parallel aligned Corpus has to be adequate for the translation domain.
- When searching such corpora the main problem is, that 1 chunk in the source has more than 1 translation in the target language and the choice is made according to the context.

Parallel aligned Corpus - Example 1-

<DOC de-news-1996-10-02-1>

<H1>

Streit um baden-wuerttembergischen Nachtragshaushalt

</H1>

Die Oppositionsfraktionen im baden wuerttembergischen Landtag haben scharfe Kritik an der Finanzpolitik der CDU/FDP-Koalition geuebt. Bei der Vorlage des zweiten Nachtragshaushalts fuer das laufende Haushaltsjahr begruessten sie zwar die strikte Begrenzung der Neuverschuldung, gespart werde aber vor allem bei den Familien und damit am falschen Fleck.Finanzminister Mayer-Vorfelder verteidigte den eingeschlagenen Sparkurs. Der Nachtragshaushalt soll rund 1,1 Mrd. DM Deckungsluecke ausgleichen, die vor allem durch Steuerausfaelle im Haushalt klaffen. Rund 800 Mio. DM sollen durch Investitions-und Sachmittelkuerzungen erbracht werden. Einsparungen im Personalbereich werden mit 130 Mio. DM beziffert.

<DOC de-news-1996-10-02-2>

.....

<DOC de-news-1996-10-02-1>

<H1>

Baden-Wurttemberg supplementary budget dispute

</H1>

The opposition parties in Baden-Wurttemberg's Landtag have strongly criticized the financial policies of the governing CDU/FDP coalition. Upon presentation of the second supplementary budget for the current budget year, they approved of the strict limitation of new borrowing,but said that savings were going to be realized in the wrong place - on the backs of families. Finance Minister Mayer Vorfelder defended the budget, saying it would equal out a shortfall of 1.1 billion marks in state finances, which was caused primarily by tax losses. Cuts to investments and materials are expected to yield 800 million marks. Savings on personnel are estimated
at 130 million Marks.

<DOC de-news-1996-10-02-2>

....

Paragraph-Alignment

Parallel aligned corpus - Example 2-

<DOC de-news-1996-10-02-1>
Streit um baden - wuerttembergischen Nachtragshaushalt

Die Oppositionsfraktionen im baden - wuerttembergischen Landtag haben scharfe Kritik an der Finanzpolitik der CDU / FDP - Koalition geuebt .

Bei der Vorlage des zweiten Nachtragshaushalts fuer das laufende Haushaltsjahr begruessten sie zwar die strikte Begrenzung der Neuverschuldung , gespart werde aber vor allem bei den Familien und damit am falschen Fleck .

Finanzminister Mayer - Vorfelder verteidigte den eingeschlagenen Sparkurs . Der Nachtragshaushalt soll rund 1,1 Mrd. DM Deckungsluecke ausgleichen , die vor allem durch Steuerausfaelle im Haushalt klaffen .

.....

<DOC de-news-1996-10-02-1>
Baden - Wurttemberg supplementary budget dispute

The opposition parties in Baden - Wurttemberg 's Landtag have strongly criticized the financial policies of the governing CDU / FDP coalition .

Upon presentation of the second supplementary budget for the current budget year , they approved of the strict limitation of new borrowing , but said that savings were going to be realized in the wrong place - - on the backs of families .

Finance Minister Mayer - Vorfelder defended the budget , saying it would equal out a shortfall of 1.1 billion marks in state finances , which was caused primarily by tax losses .

...

Sentence-Alignment

2 Sentences in the DE-Corpus correspond to 1 Sentence in the EN-Corpus

Parallel aligned Corpus - Example 3 -

Die $_1$ Oppositionsfraktionen $_2$ im $_3$ baden - wuerttembergischen $_4$ Landtag $_5$ haben scharfe $_6$ Kritik $_7$ an der Finanzpolitik $_8$ der $_9$ CDU / FDP - $_{10}$ Koalition $_{11}$ geuebt $_{12}$.

The (1) opposition parties (2) in (3) Baden - Wurttemberg 's(4) Landtag(5) have strongly (6) criticized(7+12) the financial policies(8) of (9) the governing CDU / FDP (10) coalition(11) .

Fertitlity of a source word = the number of words in the target text

e.g. *fertility(Oppositionsfraktionen) = 2*

Distortion = Source and target words do not appear in the same place e.g. Koalition und coalition

Alignment-Methods

- Manual :
  - Extreme time consuming, because for real applications the corpus has to be really big.
  - Specialists with very good knowledge in both languages are needed
- Automatic with help of statistical procedures
  - E.g. length-based methods (number of words in the source and target text has to be close one to another)
  - Difficult to identify at the word-level, because for e.g. in:

Haben Scharfe Kritik geuebt <-> *have strongly criticized*

The POS change and the semantic combinat

Lexicalization + Categorial divergence

Contents

- General Principles of Corpus-based Machine Translation
- Principles and functionality of EBMT Systems
- Adding morphological and syntactic knowledge to EBMT systems
- Adding semantic knowledge to EBMT systems

„A good translator is a lazy translator“

EBMT Sources: Theory of Translation

A new transation may use as much material as possible from old translations (produced within the same domain, time, etc.).

↓

Advantages of this approach:

- spares time

- ensures the terminological and stilistic consistency

↓

Many human translations are revisions, improvements, changes of previous translations.

Analogy principle (Nagao, 1984)

EBMT sources: cognition science

- Human translations are mostly not the result of deep linguistic analysis but more of an appropriate,
  - Division of the sentence in chunks followed by
  - Translation of the components as well as
  - Combination of these components.
- The translation of the components is done through analogy with previous existent translations.

EBMT source: MAHT

- Translators use often big databases with translation examples (Translator‘s workbenches /Translation memories).
- E.g. TRADOS - a TM-system for 12 European languages
- The system searches in the database all entries in the source language similar with the input and shows their translations
- The human translator identifies the pieces which he needs, and performs their recombination.

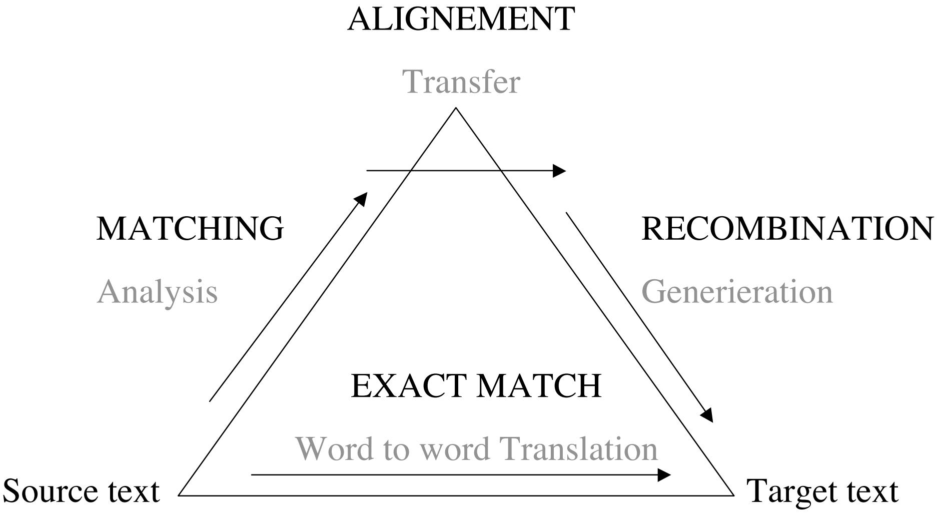
General Principles of EBMT

- A parallel Corpus is used
- Part of the input text are compared with source chunks in the corpus
- The translation of the retrieved parts are put together and form the translation.

Or

- The most similar sentences to the input in the SL corpus are retrieved (a distance is defined)
- The corresponding translations are combined to form an output

Translation pyramid for EBMT

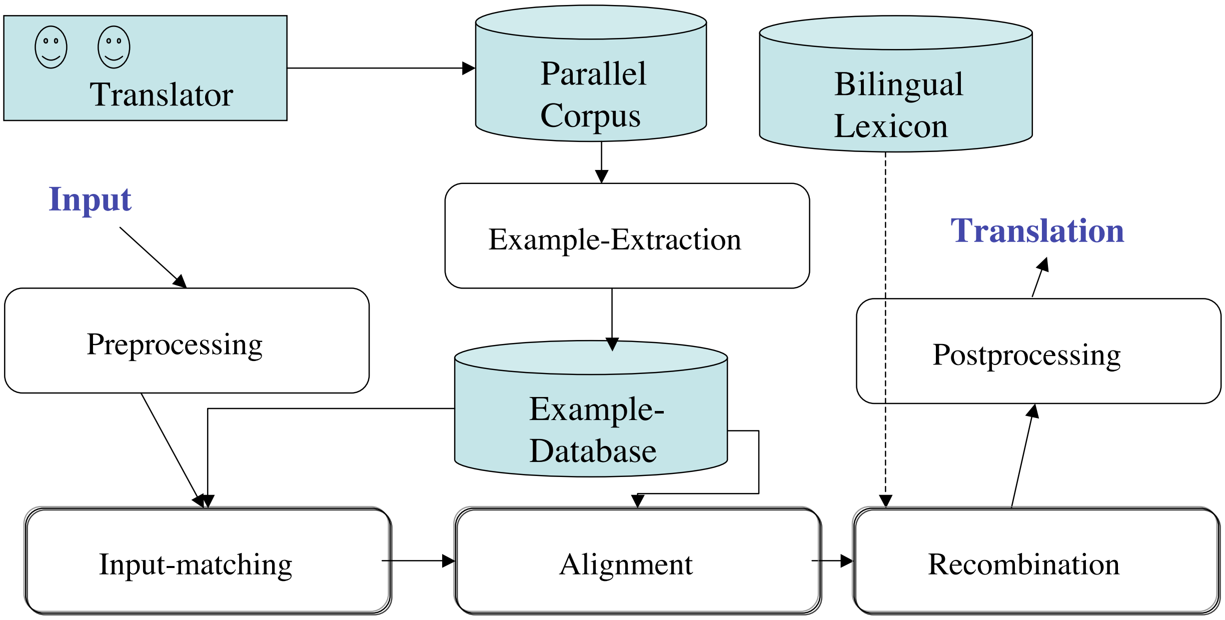


@A. Way

Functionality of an EBMT-System

- Relevant examples from a parallel corpus are extracted and saved in a database
- The input is compared with entries in the database(matching-phase).
  - Either the system looks for the identity of (parts of the) input with the database entries or
  - a distance between the input and the database entries is computed, and the database entry with the minimal distance to the input is chosen.
- Further on, in the alignment phase, the corresponding parts in the target language are retrieved (this is trivial when the whole identical input is found in the DB)
- The corresponding chunks in the target language are recombined and build the output

Architecture of an EBMT-System



Important decisions when building a database of translation examples

- Size: How many examples have to be stored?
- Length of entries: how long should be the translation examples ?
- Annotations: Do we need additional information?
- Data: What do we store (Strings, grammatical structures)?
- How do we store in order to retrieve easily

Relevant Examples?

- For a good lexical coverage:
  - a lot of domain relevant words
  - As much as possible with co-occurences (reflexiv, particle verbs, etc.)
- For a good syntactic coverage:
  - Structures containing main and relative clauses
  - Active and passive voice sentences
  - questions
  - Sentences with embedded structures, e,g attribute sentences, conjunction sentences

Lenght and Size of Examples

- The *size* of the example database varies between some hundreds and 800.000 sentences.
- The bigger the database, the better the system works
- There is no ideal *length* for the examples:
  - The longer the examples, the lower the chance for a match
  - The shorter the example the bigger the chance to have some ambiguities
- Usualy the standard *unit* for the examples is a sentence

EBMT - Example

- Input: *Ungeeigneter Kraftstoff kann zu Motorschäden führen*

- the translation database contains:

- *Starke Motorbelastung kann zu Motorschäden führen - High engine loading can cause engine damage*

-*Ungeeigneter Kraftstoff darf nicht benutzt werden.- Unsuitable fuel must not be used*

- Following chunks are identified

- *kann zu Motorschäden führen - can cause engine damage.*

-*Ungeeigneter Kraftstoff - Unsuitable fuel*

•The translation is then:

– *Unsuitable fuel can cause engine damage*

Corpus-Tagging for EBMT -1-

- It is possible to mark in the corpus words or morphemes, which delimit a clear co-text: like quantifers, conjunctions, pronouns, question markers, etc.
- E.g.

<QUANT> all uses

<QUANT> tous usages

Corpus tagging for EBMT Example Gaijin System -1-

Phrasal segmentation using Marker Hypothesis

- Psycholinguistic constraint on grammatical structure
- States that natural languages are marked for grammar by a closed set of lexemes and morphemes
- Gaijin exploits such markers as signals for beginning and end of a phrasal segment:
  - Prepositions: in, out, on, with,...
  - Determiners: the, those, a, an,....
  - Quantifiers: all, some, many,....
- Markers not considered to start a new segment if previous/next segment would consist entirely of marker words

Corpus tagging for EBMT Example Gaijin System -2-

Segment Alignement

- Possible segment correspondences between source and target are evaluated using segment length and word correspondence weights
- Bonus for having leading marker of the same category type (e.g. „with“ and „mit“)
- Many-to-one segment mappings are (partially) handled by merging contigous segments which all map to the same segment in the other language
- Non-contigous mappings are considered unusable

Corpus tagging for EBMT
Example Gaijin System -3-

Templates

- All well-formed segment mappings are converted into variables, generating a template for the translation example
- Non frequent marker words are removed from the variablized segment and retained in the template literally
- To simplify lookups, segment merging is represented in the target side only; when the source segments need to be merged the system uses a compound variable on the target side

Corpus tagging for EBMT
Example Gaijin System -4-

Template Example

E: Displays controls for coloring the extruded surfaces

G: Durch Klicken auf dieses Symbol lassen sich Optionen zum Kolorieren der extrudierten Flaechen anzeigen

Template:

E: {_A}{prep B}{det C}

G: Durch klicken auf {prep A}{prep B}{det C} anzeigen

Chunks

A: Displays Controls
dieses Symbol lassen sich Optionen

B: for coloring
zum kolorieren

C: the extrudede surfaces
der extrudierten Flaechen

How to organise the database

- There is no „best solution“
- The easiest way: all the words which exist in the database are stored and for each word a list with the id if the sentences where they appear is provided
- In the matching phase a treshold is fixed and only sentences containing at least the treshold number of words are compared with the input.

Input for Matching

- The problem is to find out, which parts of the input can be retrieved in the database
- This is done through a combination of string-based, statistical-based methods (e.g. big probability for multi-word lexemes), and help of additional linguisitc knowledge.
- String-based matching approaches:
  - Edit distance
  - Angle of similarity
  - Semantic similarity

String-based Matching

- The similarity is measured between the input string and each string in the database. Following distances are used:
  - "longest common sequence"
  - "Edit distance": how many operations (Insert, Delete, Replacement) are necessary to transform the input string into an entry in the Database
- These methods can be implemented easier through greedy algorithm, or dynamic programming

Database-Search (Alignment) -1-

- In the best case one example in the database is identical with the input
- Usually only parts of the input are found in the database
- The simplest is the organisation of the database (no indexing, no markers, no syntactic structures), the more difficult is the retrieval both of
  - the identical parts in the SL
  - Their translation equivalents

Database-Search (Alignment) -2-

- There are elaborated statistical procedures to align the segments. They are based on staistical models of the SL and TL.
- Easier: the syntactical structures in both languages (at least for some problematic chunks) are stored and links between the SL and TL structures are provided.
- Another option is to mark at least words which delimit unambigous parts of the sntence (see marker hypothesis).

Recombination

- Without grammatical structures , or at least some markers , is is very difficult,
- When syntactical structures are provided , the procedure relies on tree unification
- For strong inflected languages is usually a morphological post-processing necessary.

Contents

- General Principles of Corpus-based Machine Translation
- Principles and functionality of EBMT Systems
- Adding morphological and syntactic knowledge to EBMT systems
- Adding semantic knowledge to EBMT systems

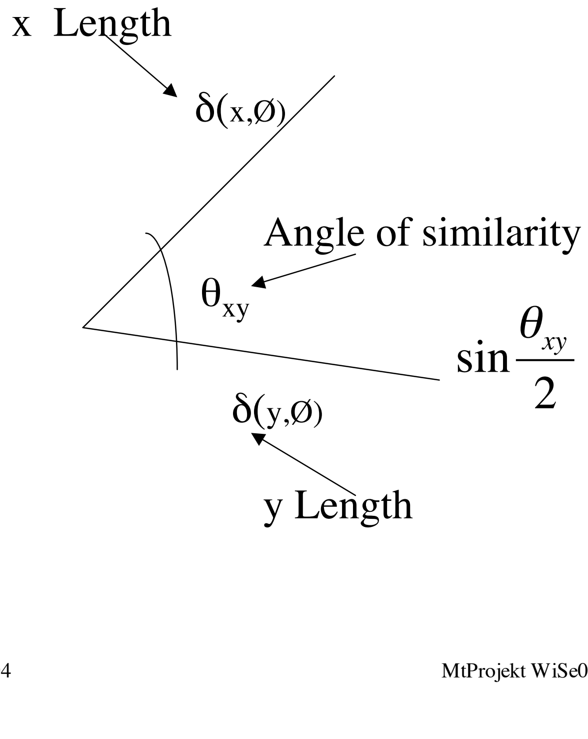
EBMT with morphological/lexical knowledge

- Use only the stems when measuring the distance between input and entries in the database
- Mark in the database words with unabigous function (e.g conjunctions)
- Whenever possible allign fixed expressions
- When measuring the Edit distance look at the PoS of the words

Word-based Matching: ”Angle of similarity” - 1 -

- A trigonometrical distance is computed.
- The distance between 2 sentences corresponds to a difference function δ.
- This difference function works similar as the string-based matching (the number of operations is calculated)
- The operations are weighted, e.g. the insertion of a comma has a smaller weight than the absence of an adjective.
- The weights are defined according to the system and the translation domain

Word-based Matching: ”Angle of similarity” - 2 -



Distance between sentence x and sentence y

$$\sin\frac{\theta_{xy}}{2} = \frac{\delta(x,y) - \left|\delta(x,\varnothing) - \delta(y,\varnothing)\right|}{2 \times \min\left\{\delta(x,\varnothing),\delta(y,\varnothing)\right\}}$$

Word-based Matching - ”Angle of similarity” Example

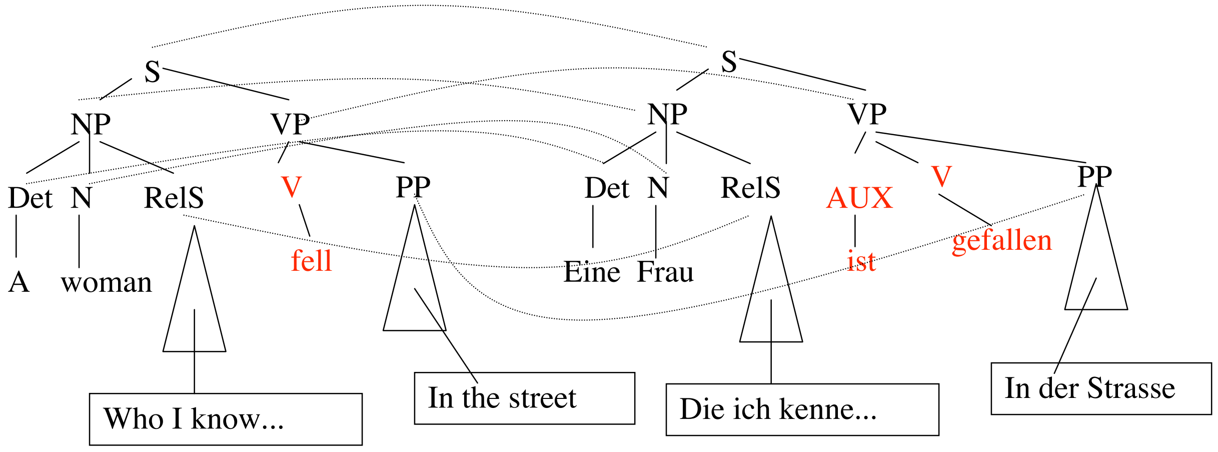
1. *Lesen Sie Seite 3 im Kapitel “Benzin”*
2. *Lesen Sie Seite 3 im Kapitel “Bremsen” und Seite 5 in Kapitel “Länderspezifische Bemerkungen”*
3. *Lesen Sie Seite 4 im Kapitel “Bremsen”.*

- String-based matching gives a closer similarity between sentence 1 and sentence 3 because they differ only by 1 word.

  However: Sentence 2 is actually a better choice as sentence 1 is contained entirely. This choice is made by the “angle distance”.

EBMT with syntactic knowledge

- The Translation patterns are not words , but syntactical structures in both languages with corresponding links
- A “semantic network“ is used additionally; in this semantical network the distances between words express semantic similarity.



Contents

- General Principles of Corpus-based Machine Translation
- Principles and functionality of EBMT Systems
- Adding morphological and syntactic knowledge to EBMT systems
- Adding semantic knowledge to EBMT systems

Word-based matching - 1 -

- For example for the following entries in the DB:
  - *Der Abstand zwischen den Kontrollen soll 2 Jahre nicht überschreiten*

    <->The interval between 2 general checks should not exceed 2 years.
  - *Der Abstand zwischen den Nebelleuchten ist x cm.*

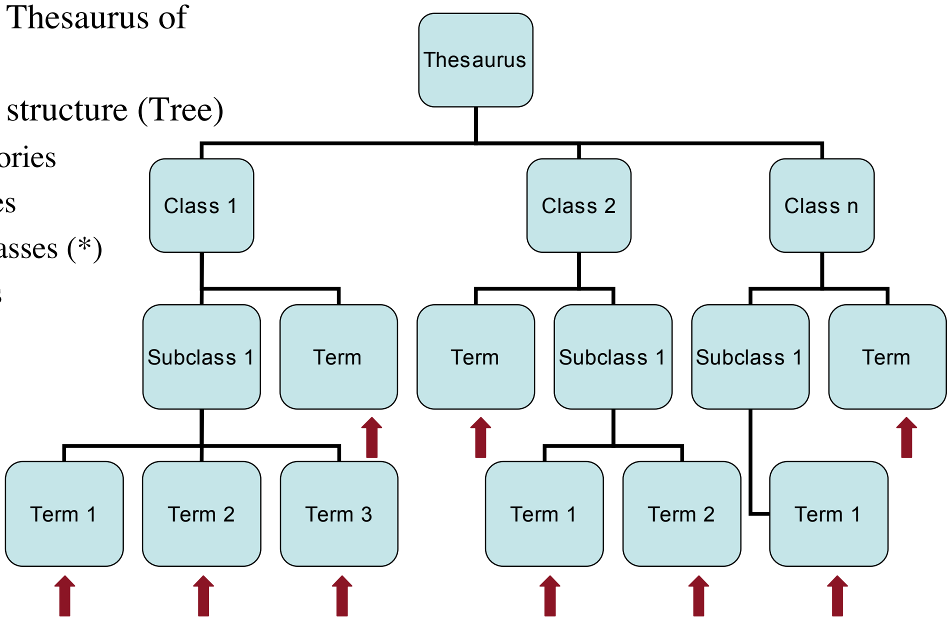
    <-> The normal distance between fog-lights is x cm.
- The input : *Wo finde ich den Abstand zwischen den Rädern?*
  - *Räder* in the semantic network is closer to *Nebelleuchten,* therefore *Abstand* is translated by *distance*,

  although the edit distance between *Räder* and *Kontrolle* is smaller than the edit distance between *Räder* and *Nebelleuchte*.

Construction of the Semantic Network (I)

- Bilingual Thesaurus of NOUNS.
- Elements structure (Tree)
  - Categories
  - Classes
  - Subclasses (*)
  - Terms

NOUNS



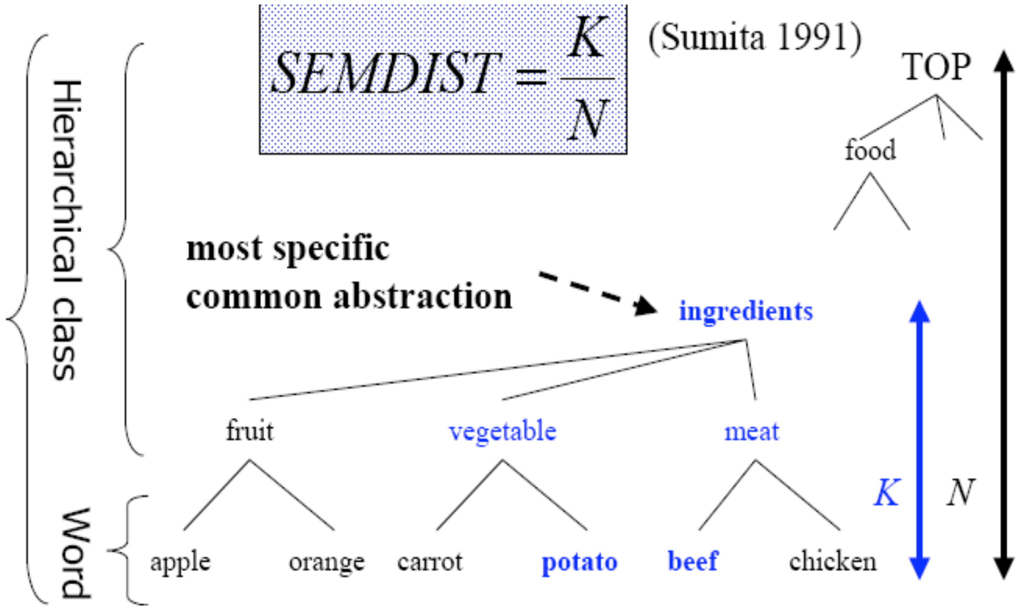
Construction of the Semantic Network (II)

- Spanish Culture
  - Entertainment
    - Fashion
    - Sports
  - Religion
  - Dietary Habits
    - Mediterranean Diet
      - Typical Food
      - Tapas
  - Art
    - Monuments
      - Mosque
      - Museum
      - Monastery
      - …
- Spanish Geography
  - Territories (“map”)
    - Autonomous Region
    - City
    - Province
    - Town
    - …
  - Geographical Quirks (“geo”)
    - Mount
    - Mountain
    - Mountain Range
    - River
    - Ocean
    - …
  - Cardinal Points

Measuring the Distance (I)

- Distance

$$dist = \frac{I + D + 2\sum semdist}{L_{input} + L_{example}}$$



$$SEMDIST = \frac{K}{N}$$

(Sumita 1991)

Measuring the Distance (II)

- Semantic Distance
  - If two words are in the same subclass -> Semantic Distance = 0. Maximal Similarity.
    - Sea – Mountain -> SD = 0
  - If they are in different categories -> Semantic Distance = 1. Completely Dissimilar.
    - Sea – Museum -> SD = 1

Measuring the Distance. Sample

- Initial sentence manipulation (lexicon):
  - INPUT: “I have seen the Alhambra of Granada”

↓

“see the monum of map”

  - CORPUS : “You will see the Mosque of Cordoba”

↓

“see the monum of map”

0 insertions 0 deletions 0 substitutions
dist = 0

Measuring the Distance. Sample

- Initial sentence manipulation (lexicon):
  - INPUT: “The autonomous region of Andalusia lies in the south of Spain”

↓

“The region of map lie in the cardinal point of map”

  - CORPUS : “The gulf of Almeria lies in the east of Andalusia”

↓

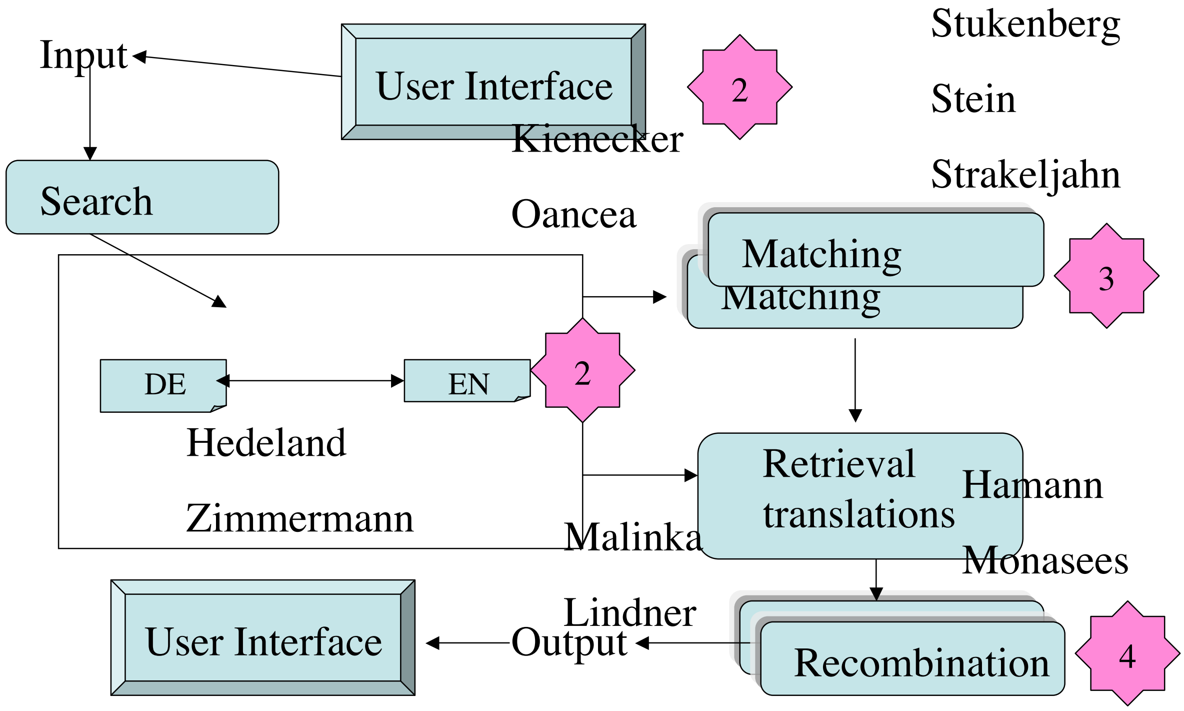
“The gulf of map lie in the cardinal point of map”

0 insertions 0 deletions 1 substitutions
semdist (region, gulf) = 0.5
dist = (0+0+2*0.5) / (11+11)

Contents

- General Principles of Corpus-based Machine Translation
- Principles and functionality of EBMT Systems
- Adding morphological and syntactic knowledge to EBMT systems
- Adding semantic knowledge to EBMT systems

Architecture of the System -Version 1



Stukenberg

Stein

Strakeljahn

Input

User Interface

2

Kienecker

Oancea

Search

Matching

Matching

3

DE

EN

2

Hedeland

Zimmermann

Retrieval translations

Hamann

Monasees

Malinka

Lindner

User Interface

Output

Recombination

4